DOCUMENT RESUME

ED 319 801 TM 015 101

AUTHOR Klockars, Alan J.; Hancock, Gregory R.
TITLE Competing Strategies for Planned Orthogonal Contrasts.
PUB DATE 90
NOTE 17p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Hypothesis Testing; Monte Carlo Methods; Research Methodology; *Simulation; *Statistical Significance
IDENTIFIERS Experimentwise Error; *Orthogonal Comparison; *Planned Comparisons

ABSTRACT

Two strategies, derived from J. P. Schaffer (1986), were compared as tests of significance for a complete set of planned orthogonal contrasts. The procedures both maintain an experimentwise error rate at or below alpha, but differ in the manner in which they test the contrast with the largest observed difference. One approach proceeds directly to the test of the contrast with the largest difference at a reduced significance level. The other is a protected procedure, first evaluating the complete null hypothesis with an omnibus "F" test, and then proceeding to test the specific hypotheses at a more liberal significance level given that the complete null hypothesis has been rejected. Monte Carlo simulation results for three and four treatment groups indicate that the relative power of the two procedures depends on the configuration of the treatment effects contained in all contrasts. Specifically, the unprotected test favors configurations with relatively small amounts of variability due to treatment effects, while the protected test has more power in cases with a relatively large amount of treatment variability. Five data tables and one figure are included. (Author/SLD)

Competing Strategies for Planned Orthogonal Contrasts

Alan J. Klockars and Gregory R. Hancock
University of Washington

Running Head: COMPETING STRATEGIES

Abstract

Two strategies, derived from Schaffer (1986), were compared as tests of significance for a complete set of planned orthogonal contrasts. The procedures both maintain experimentwise error rate at or below alpha but differ in the manner in which they test the contrast with the largest observed difference. One approach proceeds directly to the test of the contrast with the largest difference at a reduced significance level. The other is a protected procedure, first evaluating the complete null hypothesis with an omnibus F test, and then proceeding to test the specific hypotheses at a more liberal significance level given that the complete null hypothesis has been rejected. Simulation results indicate that the relative power of the two procedures depends on the configuration of the treatment effects contained in all contrasts. Specifically, the unprotected test favors configurations with relatively small amounts of variability due to treatment effects, while the protected test has more power in cases with a relatively large amount of treatment variability.

Competing Strategies for Planned Orthogonal Contrasts

How should an experimenter conduct the tests of significance associated with a 2X2 factorial design, a trend analysis, or any other design in which planned orthogonal contrasts provide the answers to the questions of interest? Should the experimenter conduct an omnibus F test and then proceed to the individual contrasts only if the omnibus test rejects the complete null hypothesis; or, should the omnibus test be bypassed, making the individual constrasts the first tests of significance conducted? In part the answers to these questions reflect researchers' position on the relative importance of control over power and Type I error. If experimenters skip the omnibus test and conduct each of the planned orthogonal contrasts at a particular per-comparison error rate (usually .05), then they will have more power (and a greater chance of a Type I error) than colleagues who use either an omnibus test as an additional control over Type I error or uses an experimentwise error rate to control Type I error. The present paper is not concerned with entering into the power versus Type I error debate. Rather, an exploration is presented of the relative power of two different strategies for conducting planned orthogonal contrasts, both of which control experimentwise Type I error for the complete null hypothesis or partial null hypotheses at a given alpha level. Thus, power differences are not purchased at the expense of control over Type I error, but rather by the configuration of the particular decision structures within each strategy.

The most common procedure for controlling the experimentwise Type I error rate is to use Bonferroni's inequality to generate per-comparison error rates. Dunn (1974) suggested conducting each of the m planned comparisons at the alpha/m level of significance; the sum of the m contrasts each conducted at this level guarantees an experimentwise error rate of no more than alpha. Following this approach, a set of k-1 planned orthogonal contrasts on k group means would involve conducting each contrast at the alpha/(k-1) level of significance. In addition, if an omnibus F test were to be conducted prior to the individual tests, the experimentwise Type I error rate would be even further reduced. This would be true whether conducting planned pairwise comparisons or planned orthogonal contrasts.

Recently, Shaffer (1986) proposed an alternative procedure for pairwise comparisons that can be applied to the testing of planned orthogonal contrasts among treatment groups. The procedure is a modification of work by Holm (1979) on applications of Bonferroni's inequality, and involves putting the test statistics T_i for all m planned comparisons in order of decreasing magnitude of absolute effect $[|T_1| \geq |T_2| \geq ... \geq |T_m|]$. In Holm's procedure, the null hypothesis for largest test

statistic H_1 is evaluated against a critical value at the alpha/m significance level. The null hypothesis corresponding to the second largest test statistic H_2 is then tested if and only if the largest comparison results in a rejected null hypothesis, and is evaluated at the alpha/(m-1) significance level. Thus, the general form of Holm's procedure is to reject hypotheses $H_1 ... H_j$, where j is the largest integer from 1 to m, such that the test statistic T_i exceeds the critical value at the alpha/(m-i+1) significance level for all i from 1 to j. Shaffer's modification of Holm's procedure involves testing each comparison at the alpha/t_i* significance level, where t_i* is the greatest number of possible true null hypotheses remaining given the rejection of the null hypotheses for all previous comparisons. In a pairwise comparison scheme, the logical implications of rejections of certain null hypotheses make the number of possible true null hypotheses remaining t_i* potentially smaller than Holm's (m-i+1), thereby increasing the power at each stage of testing by using increasingly liberal significance levels.

When applied to a complete set of planned orthogonal contrasts, the procedures of Holm and Shaffer become identical. Thus, for a set of k-1 planned orthogonal contrasts on k group means, the first contrast is evaluated at alpha/(k-1), the second at alpha/(k-2), and so on. Shaffer (1986) proved that this "modified sequentially rejective Bonferroni" (MSRB) procedure controls the experimentwise error rate below alpha for the complete null hypothesis or any pattern of true partial null hypotheses. It is also uniformly more powerful than using the simple application of Bonferroni's inequality as suggested by Dunn (1974). Because the MSRB is more powerful than Dunn's test under any configuration of treatment effects while maintaining the same control over Type I error, Dunn's approach is not considered in the present investigation.

Another approach to testing planned comparisons, also outlined in Shaffer (1986), is related to her earlier work on pairwise comparisons (Shaffer, 1979). The omnibus F test is used to evaluate the overall hypothesis that all means come from a common population. If this hypothesis is rejected the null hypothesis for the comparison whose test statistic has the greatest absolute value is evaluated at the alpha/t_i* significance level, where t_i* is the number of possible true null hypotheses given that the complete null hypothesis is false. Applying this strategy to a complete set of planned orthogonal contrasts, t_1* will be one less than the number of contrasts, or k-2, where k is the number of treatment groups. The value of t_2* will also be k-2, since rejection of the null hypothesis for the first contrast does not reduce the number of possible true null hypotheses remaining from that which was expected based upon rejection of the overall null hypothesis. The procedure continues testing the null hypothesis for each contrast with successively smaller test statistics at the alpha/(k-i) significance

level if and only if all previous null hypotheses have been rejected. This method will be labeled the "F modified sequentially rejective Bonferroni" (FMSRB) procedure.

The overall decision structures of the MSRB and FMSRB are summarized in Figure 1. It is the purpose of this paper to evaluate the relative power of the MSRB and FMSRB, and to verify control of Type I error rates. To accomplish this two series of simulations were undertaken -- the first series involved k=3 treatment groups while the second series involved k=4 treatment groups.

Insert Figure 1 about here

Simulation

k=3 treatment groups

For three treatment groups there are two orthogonal contrasts. The centers of the ten bivariate t-distributions manipulate the truth or falsehood of the null hypotheses for those contrasts, as well as the magnitude of the treatment effect given a false null hypothesis. The origin of this distribution (0,0) represents the case where both null hypotheses are true. One simulation looked at this case for an evaluation of the control over Type I error. Another case is where one contrast represents a true null hypothesis while the second contrast has a false null hypothesis. For this situation three simulations estimated the Type I error rate for the true null hypothesis and the power to detect the false null hypothesis, with the magnitude of the treatment effect built into the second contrast varied to simulate small, medium, and large treatment effects. A final case, in which both null hypotheses are false, was explored using six simulations, representing all combinations of small, medium, and large treatment effects for two contrasts. For these simulations a small treatment effect is defined as a difference whose expected value is one standard error of the difference between means away from the origin, (0,0), while medium and large treatment effects are defined as two and three standard errors from (0,0), respectively.

For this series, each replication within each simulation consisted of three groups of ten independent observations sampled from a normal distribution. Individual observations were generated by combining 24 randomly drawn numbers from the uniform distribution RANF available on Fortran IV. After transformation to a distribution with mean 50, variance 10, the observations were modified to reflect treatment effects by the addition of the appropriate constants. Ten-thousand replications were conducted for each simulation. The Type I error rates and power estimates for the

MSRB and FMSRB within a simulation were calculated for the same 10,000 replications. Each of the simulated conditions is based on different observations as a separate randomly chosen seed was selected for each.

Results and Discussion for k=3

In the simulation with both null hypotheses true, the obtained estimates of experimentwise Type I error rate are .049 for the MSRB and .046 for the the FMSRB. For the case with one true and one false null hypothesis, Table 1 presents the power estimates and Type I error rates for the three simulation. Overall, the power of the MSRB is greater than that of the FMSRB for this configuration. For small treatment effects the difference is less than 1%, for large effects slightly less than 2%, while for medium effects the difference is 2.2%. The similarity of the result for the large and medium treatment effects conditions reflects a less extreme definition of large effects (approximately 75% chance of rejecting the null hypothesis) than of small effects (approximately 10%). In all configurations with true null hypotheses, control over Type I error was maintained.

Insert Table 1 about here

For the case of two false null hypotheses, the results of the six simulation configurations are presented in Table 2. Four measures of power are reported: probability of rejecting contrast 1, probability of rejecting contrast 2, probability of rejecting either of the contrasts, and probability of rejecting both contrasts. All represent power estimates since, in these simulations, both null hypotheses are false. The latter two measures correspond closely to any-pair power and all-pair power as used by Ramsey (1978).

Insert Table 2 about here

In these simulations the power of the FMSRB is slightly greater than for the MSRB on all contrast configurations except [Large, Small]. When both contrasts contribute systematically to the Mean Square Between Treatments, the omnibus F test is more likely to reject the complete null hypothesis, with the FMSRB then proceeding to the test of the two specific hypotheses. At that point the critical value required of the contrast with the greater t value would be 2.365 ($t_{.025}$) for the MSRB, while for the FMSRB the critical value would be 2.052 ($t_{.05}$). The smaller contrast would be evaluated against a critical value of 2.052 ($t_{.05}$) for both procedures. The "Any Contrast" column in

Table 2 reflects the largest differences in power between the procedures for the largest test statistic, since the tests of the smaller contrast are identical. These differences range from less than 1% to greater than 5%, with the magnitude of the difference being greater when all contrasts have moderate and comparable treatment effects.

k=4 treatment groups

The second series of simulation used four treatment groups each with n=10 randomly generated scores. As before the scores were generated by summing 24 randomly chosen numbers from the RANF uniform distribution. A complete set of three orthogonal contrasts was defined on the four groups. Two contrasts were of the form $t=(\bar{X}_i-\bar{X}_j)/\sqrt{(2MS_w/n)}$. The first compared groups 1 and 2 while the second compared groups 3 and 4. The remaining contrast was of the form $t=[(\bar{X}_1+\bar{X}_2)-(\bar{X}_3+\bar{X}_4)]/\sqrt{(4MS_w/n)}$.

As before the treatment effect conditions were achieved by separating the means by zero, one, two, and three standard errors for the null, small, medium, and large treatment effects, respectively. The Type I error rate and power for the 20 unique configurations of these four effects were estimated by simulations. One simulation reflected the completely true null hypothesis. Three simulations involved two true partial null hypotheses, while six involved one true partial null hypothesis. The remaining ten simulations reflected situations where all three contrasts were of false null hypotheses.

Results and Discussion for k=4

The experimentwise Type I error rate for the simulation with all three null hypotheses true was .047 for the MSRB and .037 for the FMSRB. For the case with two true and one false null hypothesis, the observed power and experimentwise Type I error rates are presented in Table 3. In all three such simulations the MSRB was more likely to detect the difference than was the FMSRB. The difference exceeds 4-5% in those simulations with moderate and large treatment effects. In all configurations control over Type I error was maintained.

Insert Table 3 about here

For the case where two null hypotheses were false and one was true, six simulations estimated the power and experimentwise Type I error rates. These results are presented in Table 4, demonstrating that the MSRB tends to be more powerful when there is little systematic variance within the set of means. As more variability is introduced in medium and large treatment effect conditions the FMSRB becomes slightly more powerful than MSRB. Both procedures provide

conservative control over experimentwise Type I error rate.

Insert Table 4 about here

Table 5 presents the results of the simulations for the condition where all three contrasts have false null hypotheses. The first pair of columns presents the any-pair power associated with detecting one or more of the false null hypotheses. The middle pair of columns presents the probability of detecting two or more false null hypotheses, and the last two columns present the probability of correctly detecting all three false null hypotheses. The FMSRB is generally more powerful than the MSRB for detecting the first contrast, as long as overall there is sufficient systematic variation in the group means to reject the omnibus test. The two simulations where the reverse was true are [Small, Small, Small] and [Large, Small, Small], both of which include several groups with small treatment effects. When attention is directed to detecting more than one of the treatment effects, the MSRB and FMSRB have trivial differences.

Insert Table 5 about here

Conclusions

The Monte Carlo results for both three and four treatment groups support the following general conclusions. First, both procedures provide adequate control over experimentwise Type I error whether there is a complete or partial true null hypothesis. In no instance did an estimate of Type I error for any configuration of treatment effects exceed the alpha level chosen as the maximum experimentwise error rate. In most instances the control over Type I error was quite conservative. Second, when little overall systematic treatment variance is present, the FMSRB has less power than the MSRB. But, as more systematic treatment variance is introduced either by more or larger effects, the power of the FMSRB exceeds that of MSRB. And third, the difference between the procedures is most clearly seen on the first contrast evaluated. It is on this contrast that there is a difference in the critical values required for significance; after this, both procedures use the same critical values at each remaining stage of testing.

While the magnitude of the differences in are small, the researcher can achieve increased power by selection of the appropriate decision structure. Where only one contrast is of importance the

experimenter would be best served by using the MSRB; however, where two or more contrasts are likely to contribute systematic variance to the overall F ratio, the experimenter will achieve greater power by using the FMSRB.

Two questions of generalizability are of concern with the present findings. The first concerns whether similar results would hold had a different set of orthogonal contrasts been explored. Power differences between contrasts are a function of the magnitude of the treatment effect and the standard error. To standardize the treatmen effect the current study imposed treatment effects in multiples of the appropriate standard error. Thus, the differences due to the number of groups involved in the contrast were eliminated since these differences would be reflected in the size of the standard errors.

The second concern is the generalizability of the findings to more than four treatment conditions. The differences between the two strategies are almost exclusively reflected in the evaluation of the contrast with the largest treatment effect. The critical value for this contrast will differ for the two strategies with the t-value required by FMSRB smaller than by MSRB regardless of the number of treatment groups involved. Likewise, regardless of the number of treatment groups involved the probability that the overall null hypothesis will be rejected will increase when several contrasts contribute systematic variance rather than just a single contrast. Thus, the same conclusions would be reached concerning the relative power of the two strategies regardless of the number of groups. These conclusions are that when few contrasts contribute systematic variance the omnibus F test would result in a number of incorrectly retained null hypotheses. This would more than counter any reduction in the t value for the largest contrast, and hence would result in more power with the MSRB. However, when several contrasts contribute systematic variance the complete null hypothesis is likely to be rejected and increased power will be achieved by the FMSRB due to the lower critical value for the contrast with the largest treatment effect.

Would an experimenter know enough about the treatment effects to capitalize on the differential power of the two strategies? While this information may not always be available, it is similar to that needed to conduct any power analysis to decide on an appropriate sample size. Where the experimenter is uncertain, a careful review of the literature may provide the required information.

References

Dunn, O. J. (1974). On multiple tests and confidence intervals. Communications in Statistics, 3, 101-103.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. Journal of the American Statistical Association, 73, 479-487.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range procedure. Journal of Educational Statistics, 4, 14-23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Table 1

Power and experimentwise Type I error rate when one contrast null hypothesis is true and one is false.

Treatment Effect	Observed Power		Type I error	
	FMSRB	MSRB	FMSRB	MSRB
Small (S)	.102	.106	.033	.028
Medium (M)	.364	.386	.041	.035
Large (L)	.731	.749	.050	.045

Table 2

Power when both contrast null hypotheses are false.

Contrast Effect		Observed Power							
		Contrast 1		Contrast 2		Any Contrast		Both Contrasts	
1	2	FMSRB	MSRB	FMSRB	MSRB	FMSRB	MSRB	FMSRB	MSRB
S	S	.117	.112	.112	.106	.199	.191	.030	.028
M	S	.400	.393	.147	.127	.458	.438	.089	.082
L	S	.758	.761	.159	.149	.778	.776	.138	.134
M	M	.465	.432	.458	.424	.665	.612	.258	.244
L	M	.803	.774	.480	.457	.870	.828	.414	.403
L	L	.833	.814	.827	.808	.956	.927	.703	.695

Table 3

Power and experimentwise Type I error rate when one contrast null hypothesis is false and two are true.

Treatment Effect of False H_0	Observed Power		Type I Error	
	FMSRB	MSRB	FMSRB	MSRB
Small (S)	.063	.073	.032	.031
Medium (M)	.284	.326	.039	.037
Large (L)	.649	.705	.046	.044

Table 4

Power and experimentwise Type I error rate when two contrast null hypotheses are false and one is true.

	Observed Power									
Effects	Contrast 1		Contrast 2		Any Contrast		All Contrasts		Type I error	
	FMSRB	MSRB	FMSRB	MSRB	FMSRB	MSRB	FMSRB	MSRB	FMSRB	MSRB
SS[a]	.076	.080	.069	.072	.135	.144	.010	.009	.018	.018
MS	.305	.326	.088	.084	.350	.368	.042	.042	.026	.023
MM	.352	.340	.352	.343	.550	.532	.154	.150	.028	.026
LS	.662	.697	.100	.096	.681	.713	.081	.080	.025	.024
LM	.716	.712	.378	.365	.798	.785	.296	.292	.033	.032
LL	.746	.735	.741	.731	.919	.900	.568	.566	.037	.036

[a] These symbols refer to the relative magnitude of the treatment effects contained in the first and second contrasts, respectively.

Table 5

Power when all three contrast null hypotheses are false.

Effects	Observed Power					
	One or more contrasts		Two or more contrasts		All three contrasts	
	FMSRB	MSRB	FMSRB	MSRB	FMSRB	MSRB
SSS[a]	.199	.200	.029	.028	.003	.003
MSS	.421	.412	.080	.078	.012	.012
MMS	.587	.556	.195	.191	.040	.039
MMM	.699	.649	.320	.315	.116	.114
LSS	.726	.730	.143	.140	.019	.019
LMS	.810	.789	.348	.344	.064	.064
LMM	.861	.827	.486	.482	.193	.192
LLS	.918	.897	.605	.601	.112	.111
LLM	.929	.906	.686	.682	.321	.319
LLL	.956	.942	.830	.827	.564	.563

[a] These symbols refer to the relative magnitude of the treatment effects contained in the first, second, and third contrasts, respectively.

Figure 1
Decision structures for the MSRB and the FMSRB.

For a complete set of k-1 orthogonal contrasts on k groups means, with test statistics ranked in descending the decision structure for both procedures is presented below. Note that in both procedures, advancing to the next stage of testing is contingent upon rejection of all previous null hypotheses.

MSRB	FMSRB
	1. Test complete null hypothesis at α significance level.
1. Test contrast with largest test statistic at $\alpha/(k-1)$ significance level.	2. Test contrast with largest test statistic at $\alpha/(k-2)$ significance level.
2. Test contrast with next largest test statistic at $\alpha/(k-2)$ significance level.	3. Test contrast with next largest test statistic at $\alpha/(k-2)$ significance level.
3. Test contrast with next largest test statistic at $\alpha/(k-3)$ significance level.	4. Test contrast with next largest test statistic at $\alpha/(k-3)$ significance level.
⋮	⋮
K-2. Test contrast with smallest test statistic at $\alpha/[k-(k-1)]$ (i.e. α) significance level.	K-1. Test contrast with smallest test statistic at $\alpha/[k-(k-1)]$ (i.e. α) significance level.